

Weekly Report

October 29, 2017

1 Work

This week, we studied the highway record data (20+ GB) and imported the data into Mysql database (13.8 GB) for organization. The data includes entry and exit data during 20160801 to 20160830. However, the time range of hdvd data (vehicle License plate recognition record) is from 20170801 to 20170830.

I prepared a paper report, SkyLens: Visual Analysis of Skyline on Multi-dimensional Data, for group meeting. I have reported this paper at vagblog.

For dimension reduction project, we are refactoring the code into three part, including data module, knn graph module, and embedding module for further comparison between algorithms.

1.1 工作进度

Table 1: 工作进度

TASK	PROGRESS	DATE
dimension reduction	refactor the code	11.30
hihgway project	import data into database	12.10
*2Vec survey	collect papers	12.30

2 IEDA

2.1 动态词嵌入

这个想法是直接将目前机器学习领域动态词嵌入的方法引入，加上可视化的方法分析单词词性的变化。虽然没有非常明确想要完成的任务，但我认为可以分析

研究的属性包括单词频率，相似度，时间等三个维度，以下几点非常有意思：

- 探索一个单词词性的变化（如亚马逊从一个森林相关的单词转变为科技单词）
- 两个单词（或者一对多）相似度的变化（如总统和奥巴马，在奥巴马当选总统的那几年相似度最高）
- 多个单词相似度的变化（可能一个新概念的兴起，导致原来一些不相干的单词的相似度不断提高，这个目前没有找到案例）

这个项目可以看作是对高维数据（ ~ 100 维）的探索分析，利用相似度查看他们在高维空间的关系。

2.2 通过动态词嵌入对动态网络进行布局

动态网络布局中需要考虑的一点是不同时间之间节点的位置不能变化太大，否则不利于观察分析。类似于node2vec等方法是将网络嵌入到低维空间，但是没有考虑动态网络的时间信息。因此，目前有以下几点需要克服的难点：

- 拓展基于node2vec的图嵌入方法，提出新的图布局算法，支持动态网络的布局
- 算法使得不同时间片段嵌入的低维空间要相近，减少节点位置移动。
- 算法支持streaming的网络变化（即，一开始不知道未来节点的变化情况）
- 算法支持节点属性相似性（如果节点有属性的话，嵌入的时候考虑一下属性上的相似性）

3 Paper Reading

3.1 Learning Word Relatedness over Time

为了探究单词之间关系的动态演变，作者对每一年的文本都进行单独的训练，然后计算两个单词在不同年份之间的相似性。

3.2 Temporal Word Analogies: Identifying Lexical Replacement with Diachronic Word Embeddings ACL2017

由于对不同年份的单词嵌入都是分离的，很难比较不同年份之间单词的关系（上文只是计算同一个年份的相似度）。作者使用一个全局的线性回归将不同年份计算的嵌入空间对应起来，从而比较不同年份的单词类比。

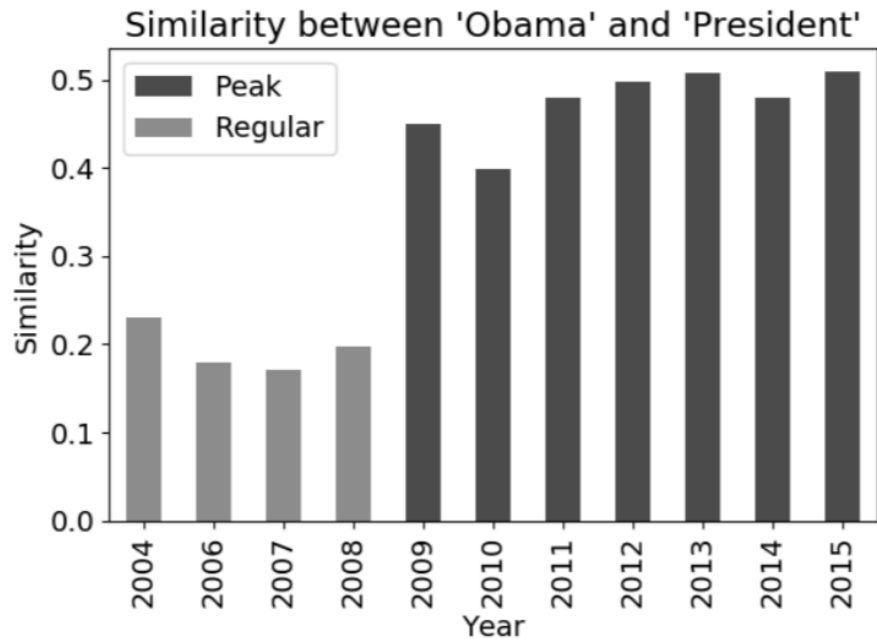


Figure 1: 1

1987	reagan	koch	soviet	iran.contra	navratilova	yuppie	walkman
1988	reagan	koch	soviet	iran.contra	sabatini	yuppie	tape_deck
1989	bush	koch	soviet	iran.contra	navratilova	yuppie	walkman
1990	bush	dinkins	soviet	iran.contra	navratilova	yuppie	headphones
1991	bush	dinkins	soviet	iran.contra	navratilova	yuppie	cassette_player
1992	bush	dinkins	russian	iran.contra	sabatini	yuppie	walkman
1993	clinton	dinkins	russian	iran.contra	navratilova	yuppie	cd_player
1994	clinton	mr_giuliani	russian	iran.contra	sanchez_vicario	yuppie	walkman
1995	clinton	giuliani	russian	white.house	graf	yuppie	cassette_player
1996	clinton	giuliani	russian	whitewater	graf	yuppie	walkman
1997	clinton	giuliani	russian	iran.contra	hingis	yuppie	headphones
1998	clinton	giuliani	russian	lewinsky	hingis	yuppie	headphones
1999	clinton	mayor_giuliani	russian	white.house	hingis	yuppie	buttons
2000	clinton	giuliani	russian	white.house	hingis	yuppie	headset
2001	bush	giuliani	russian	iran.contra	capriati	yuppie	headset
2002	bush	bloomberg	russian	white.house	hingis	gen.x	mp3_player
2003	bush	bloomberg	russian	white.house	agassi	hipsters	walkman
2004	bush	bloomberg	north.korean	iran.contra	federer	gen.x	headphones
2005	bush	bloomberg	north.korean	white.house	roddick	geek	ear_buds
2006	bush	bloomberg	iranian	white.house	hingis	teen	headset
2007	bush	bloomberg	iranian	capitol_hill	federer	dads	ipod

Figure 2: 2

3.3 Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change ACL2016

本文也是探索单词语义含义的变化，对于对齐不同年份的嵌入空间，使用了 orthogonal Procrustes 方法。

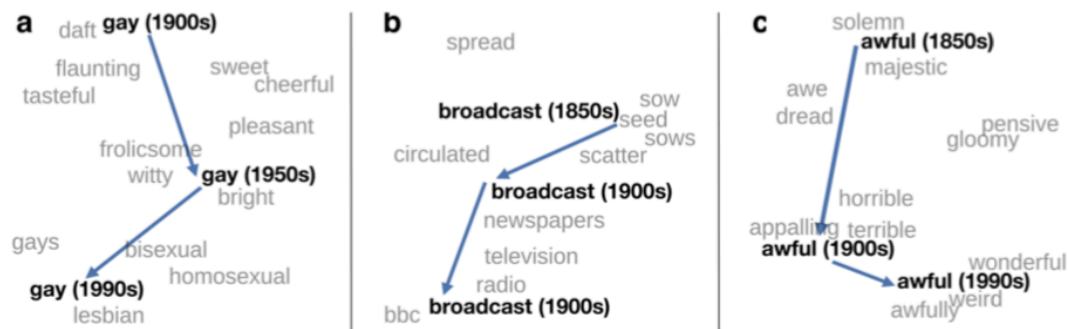


Figure 3: 3

3.4 Dynamic Word Embeddings ICML2017

本文也是处理不同年份嵌入空间不同的问题，与前面不同的是，本文提出了基于贝叶斯概率的模型，将时态变化直接由模型训练过程中建模，而不是训练好之后再处理。

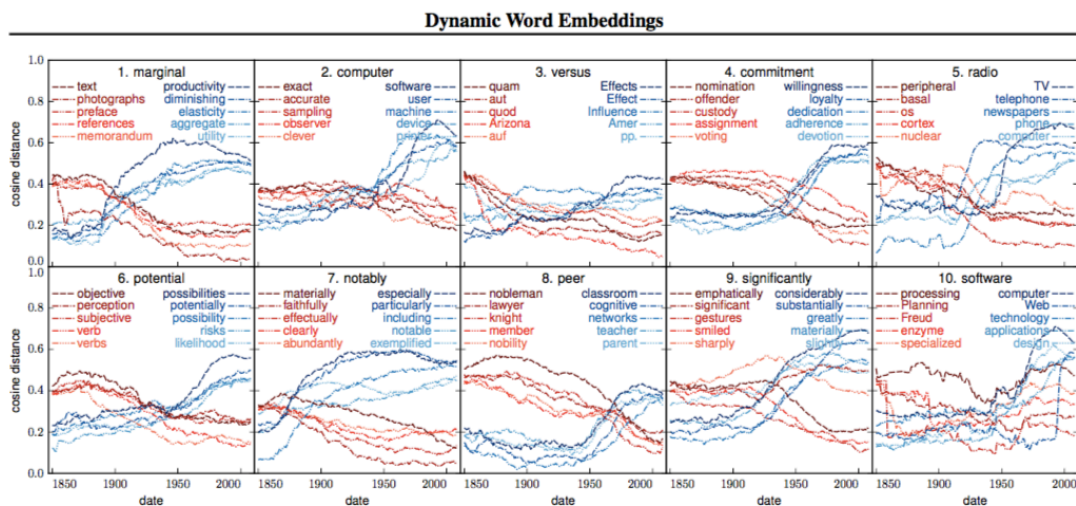


Figure 4: 4